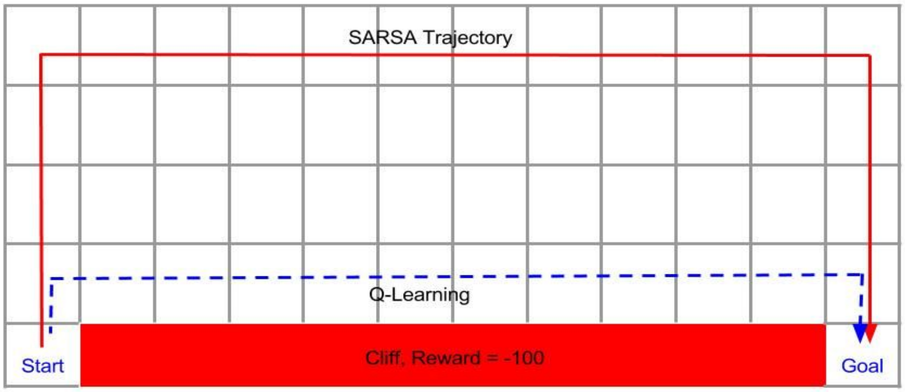
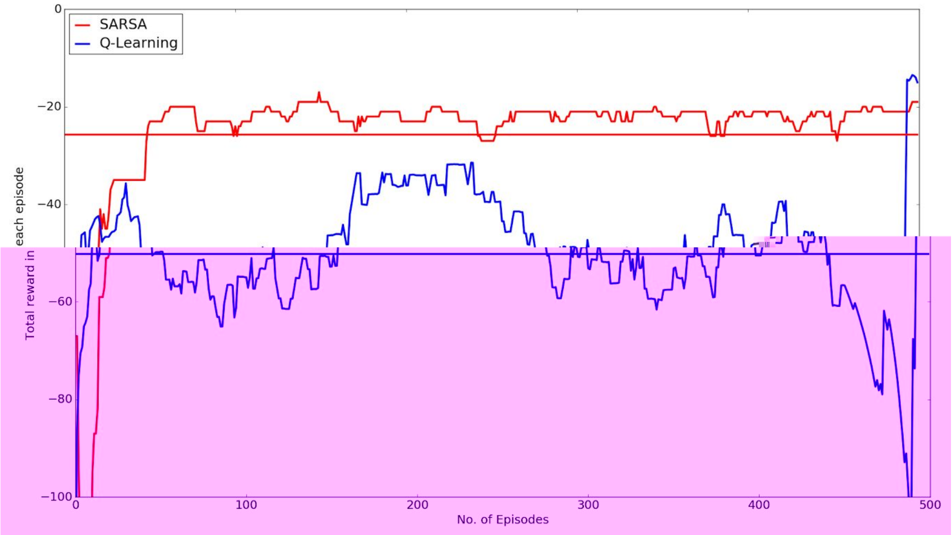
Assignment-2:Reinforcement Learning
Submitted by: Ajinkya Jain

Example 6.6: Cliff Walking: This gridworld example compares Sarsa and Q-learning, highlighting the difference between on-policy (Sarsa) and off-policy (Q-learning) methods. Consider the gridworld shown in the upper part of Figure. This is a standard undiscounted, episodic task, with start and goal states, and the usual actions causing movement up, down, right, and left. Reward is −1 on all transitions except those into the region marked \The Cliff." Stepping into this region incurs a reward of −100 and sends the agent instantly back to the
start.





Hypothesis:
In the figure shown above, the path taken by the Q-Learning algorithm is the optimal path, while the path taken by the SARSA algorithm is the safest path. The results are obtained with an ε -greedy policy having ε = 0.1. The dependence of the performance of the two learning methods on the exploration factor can be examined empirically.

The present study hypothesize that:

- *As Q-Learning is an off-policy learning method, the trajectory learned by it should not be affected by the changes in the exploration factor. However, as SARSA is an on-policy method, it should take more “safer” trajectories as exploration factor increases.*
- *And such a behavior of SARSA algorithm, should be independent of the size of grid world.*

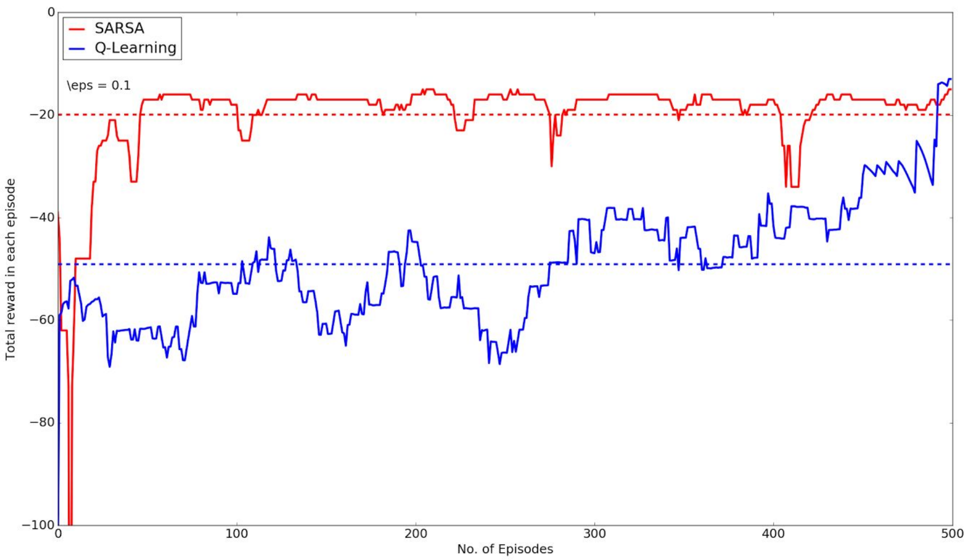
Parameters:

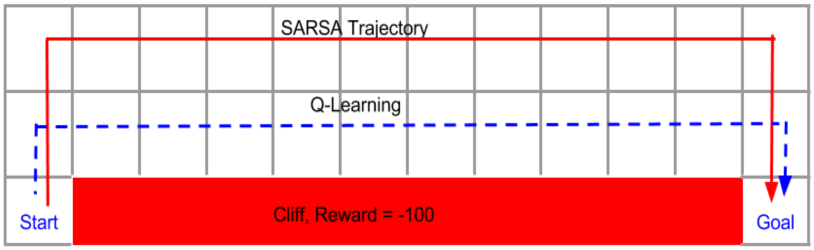
- Start state = [0,0]
- Action Set: {Right, Left, Up, Down}
- Reward Function:
  - -100 if encountered cliff, (reset to the start state)
  - -1 for each time step
- Update step size, $\alpha = 0.5$
- Discount factor, $\gamma = 1.0$

Analysis:

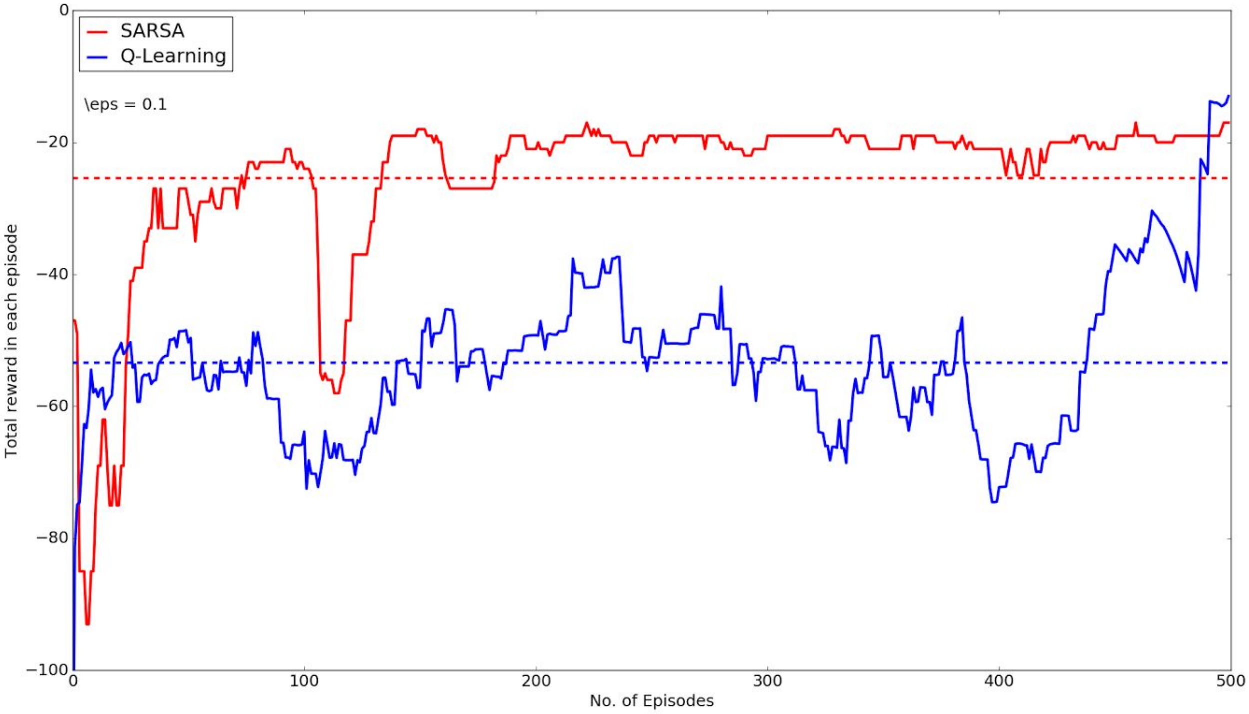
The proposed hypothesis is dependent on the two parameters, the exploration factor ε and the size of the grid world. To verify the proposed hypothesis, at first the exploration factor was kept constant at $\varepsilon = 0.1$ and the size of the grid world was varied. The following results shows the performance of the two algorithms with varying grid sizes.

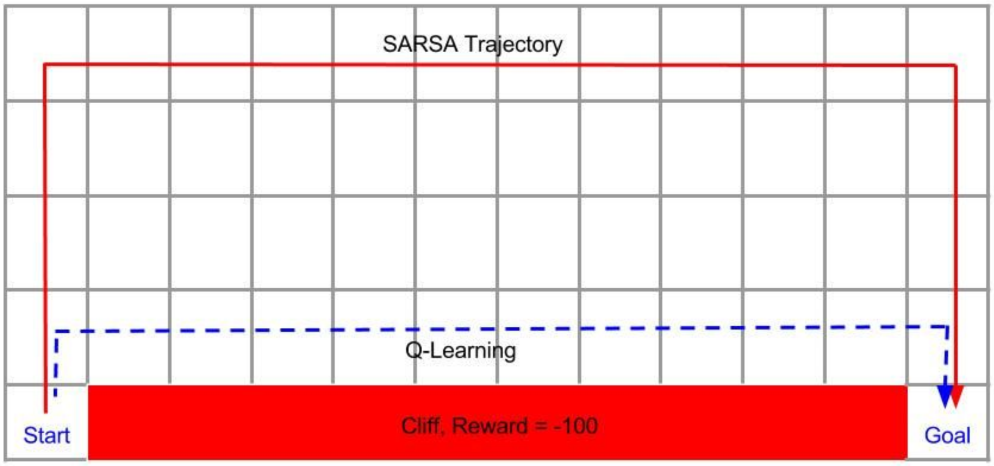
- **Changing the grid size by Breadth:**
  - Grid Size [3X12]



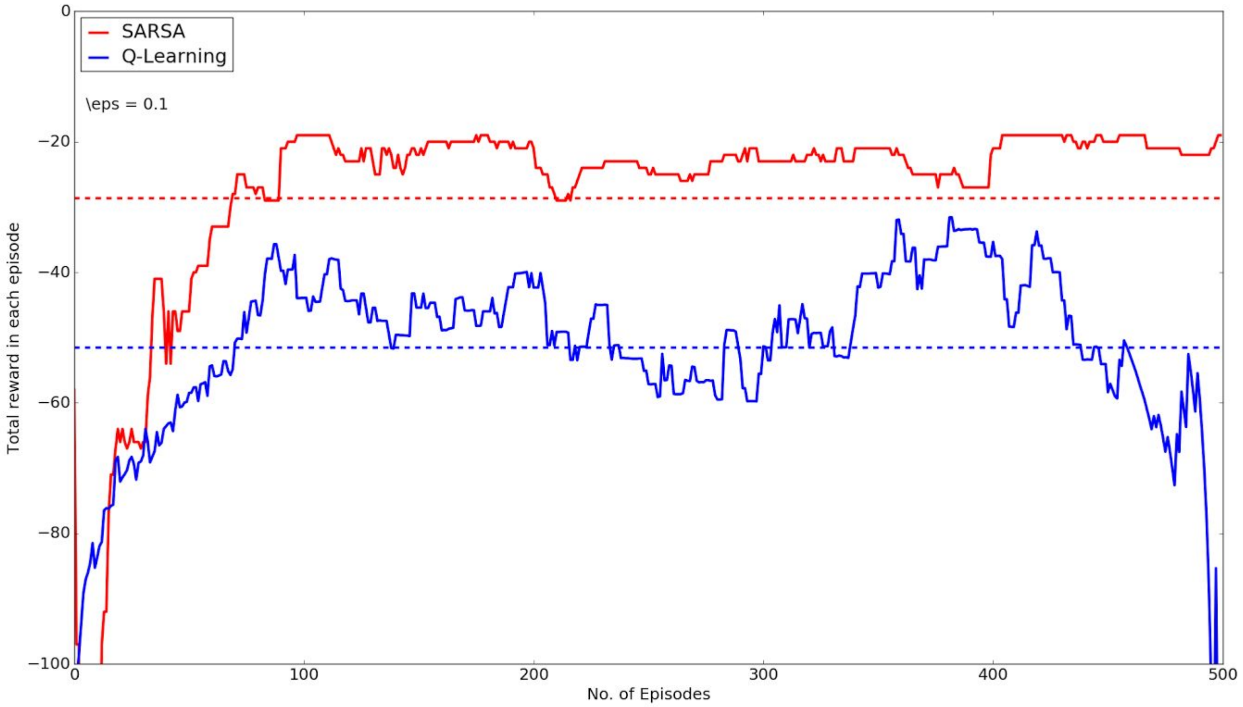


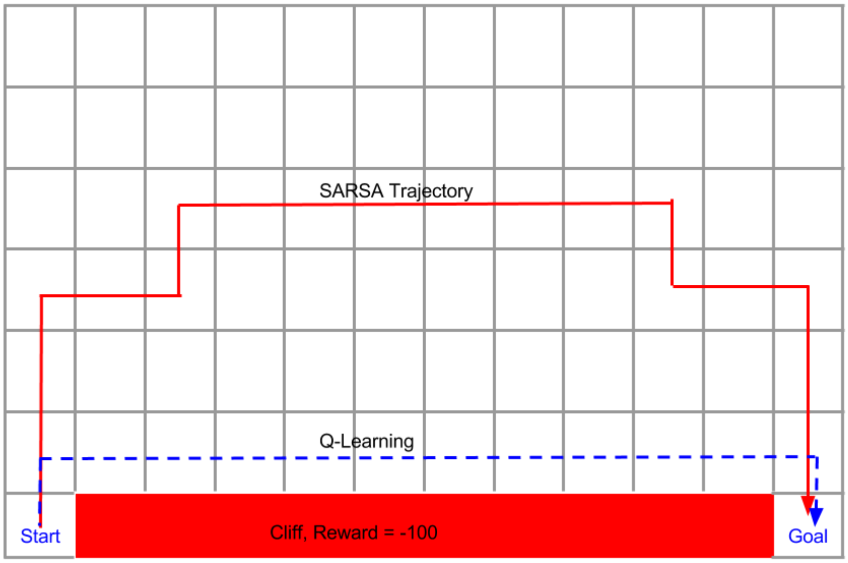
- ○ Grid Size [5X12]



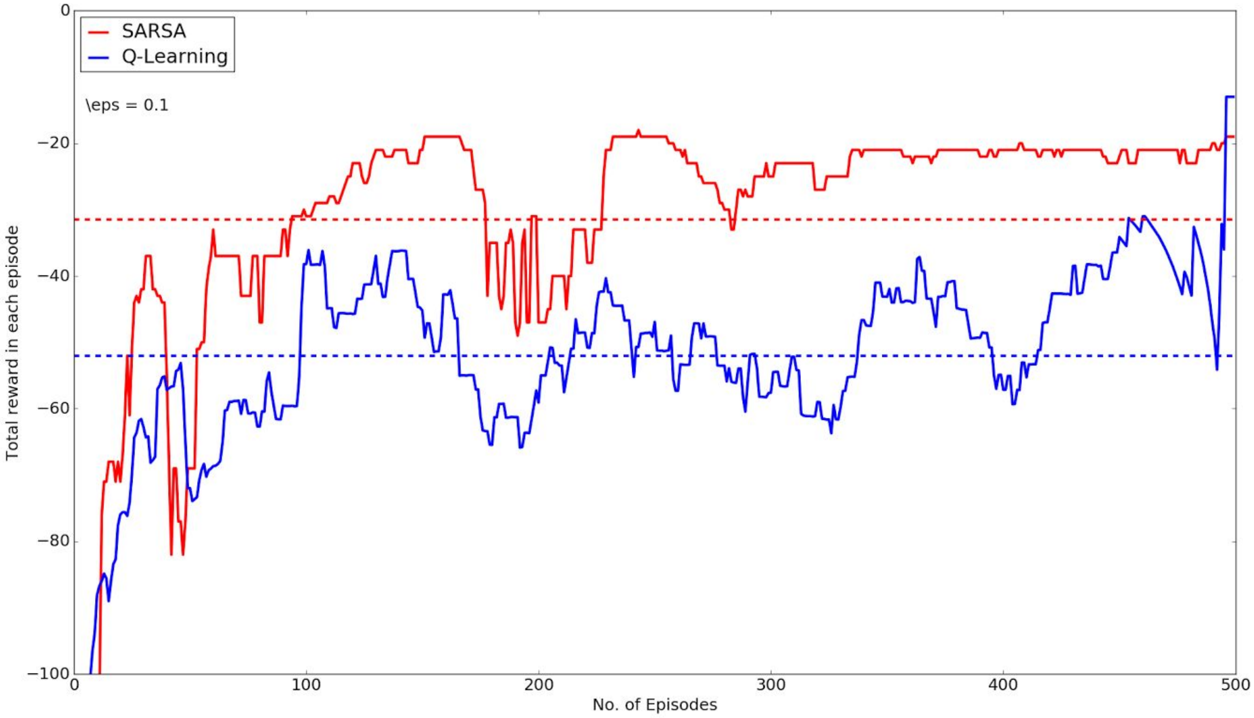


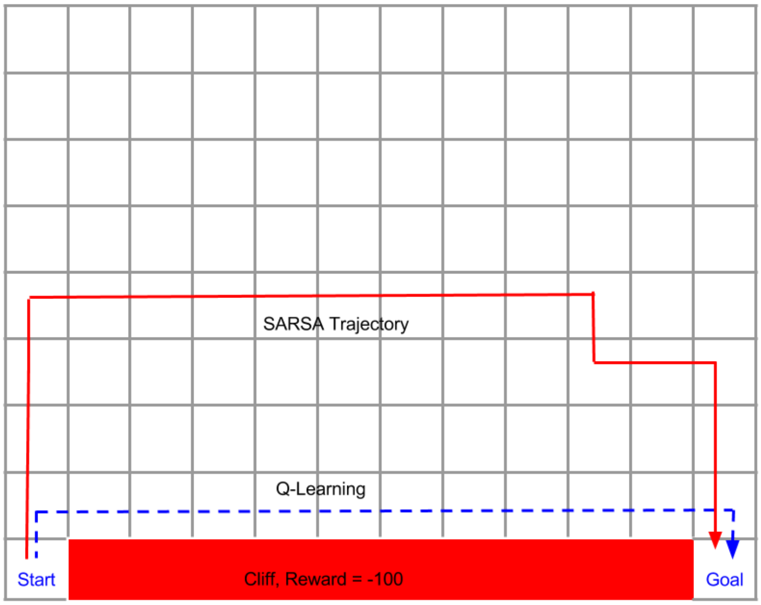
○ Grid Size [7X12]



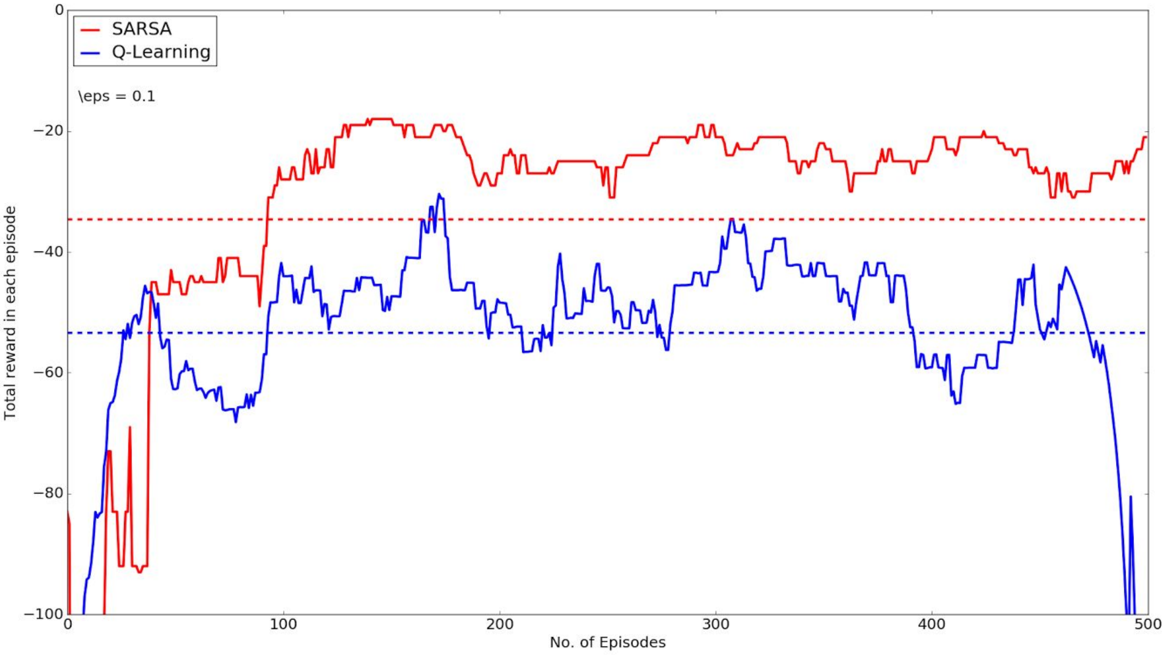


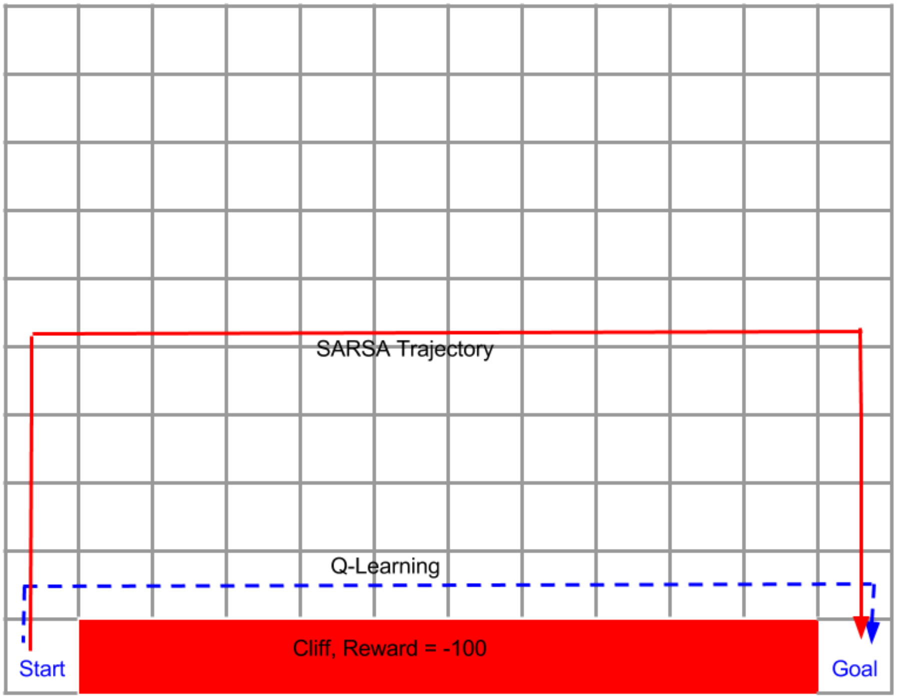
- Grid Size [9X12]





- Grid Size [10X12]

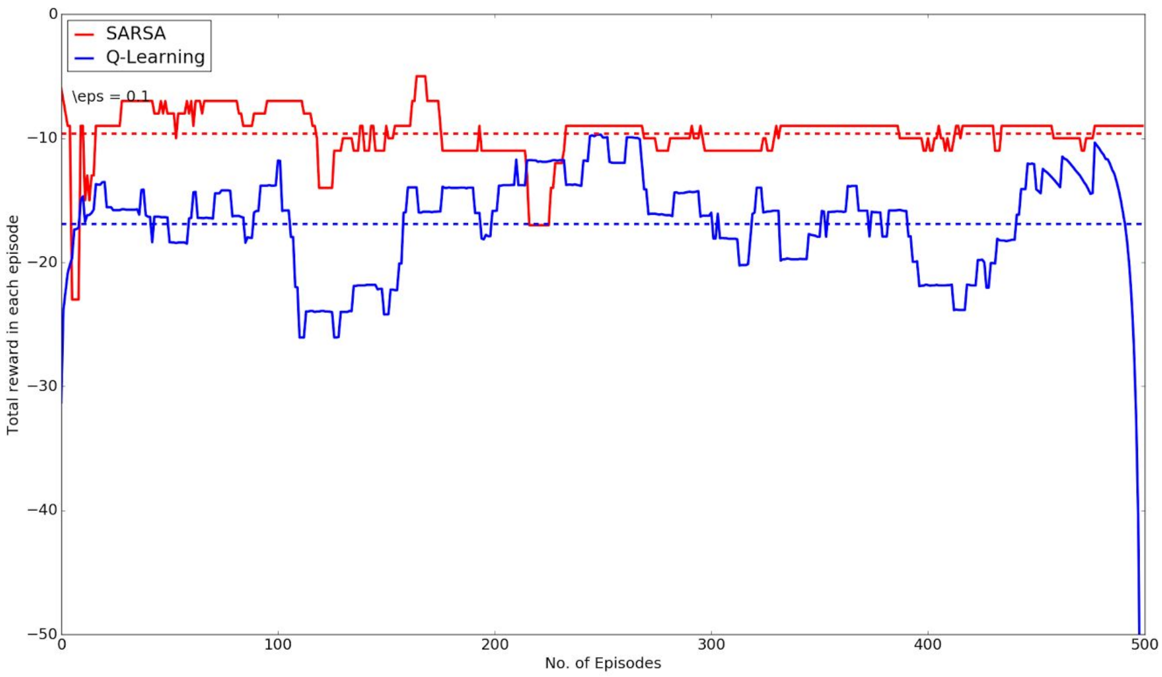


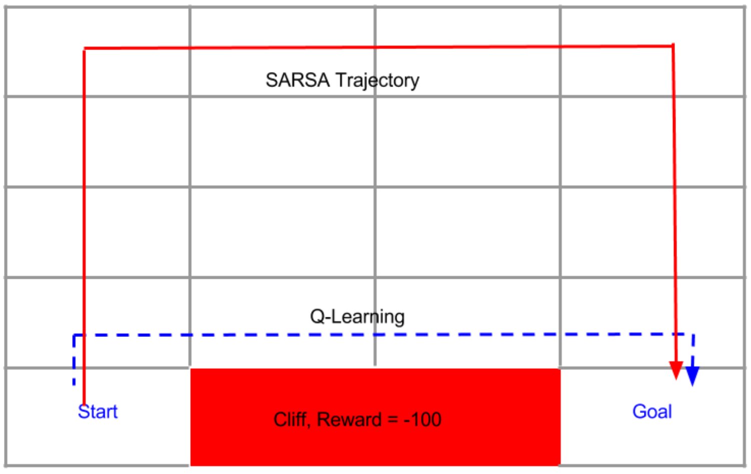


It can be seen from the results that if we keep the exploration parameter constant and vary the breadth of the grid world, the safest trajectory for the SARSA algorithm remains practically at the same offset from the cliff. Hence, *it can deduced that the given an exploration factor and a grid world of sufficient breadth, the maximum safe distance considered by the SARSA algorithm remains constant and is independent of the increase in the breadth of grid world thereupon.*

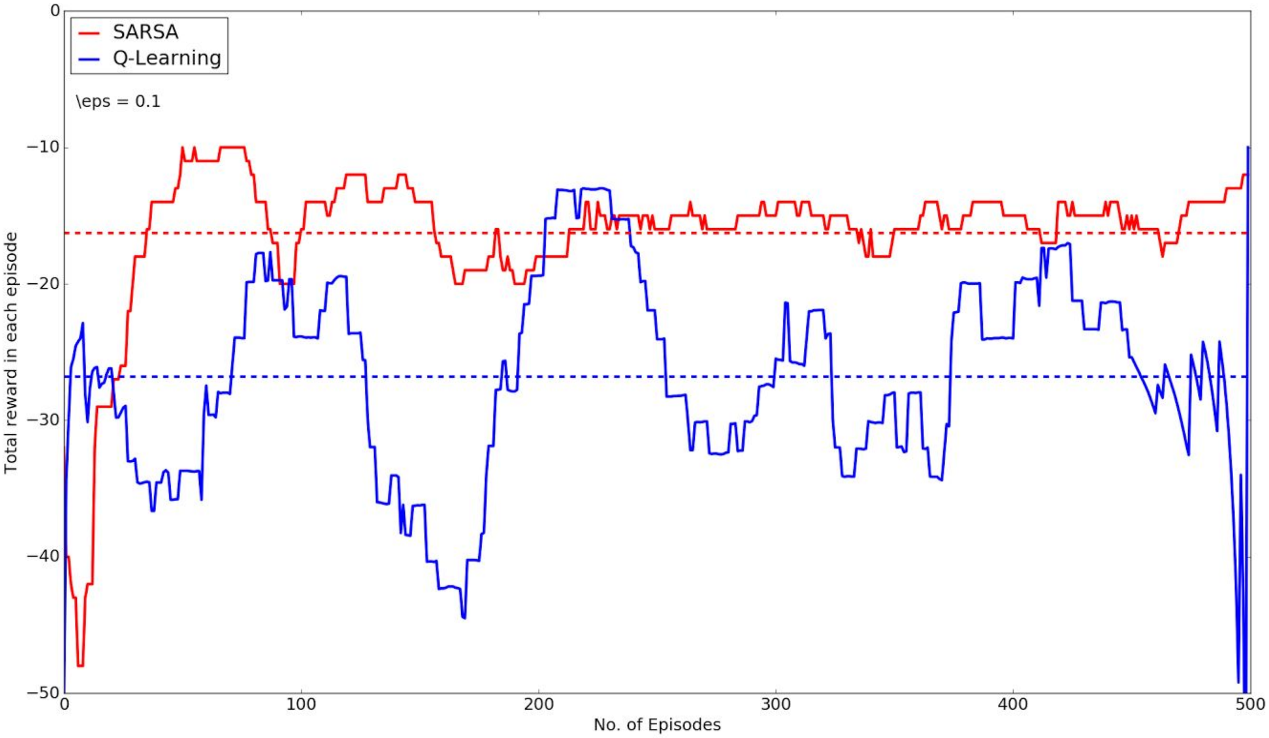
The grid can be varied along its length dimension too. Based on the aforementioned inference, the grid world for these set of experiments is considered to be of breadth 5.

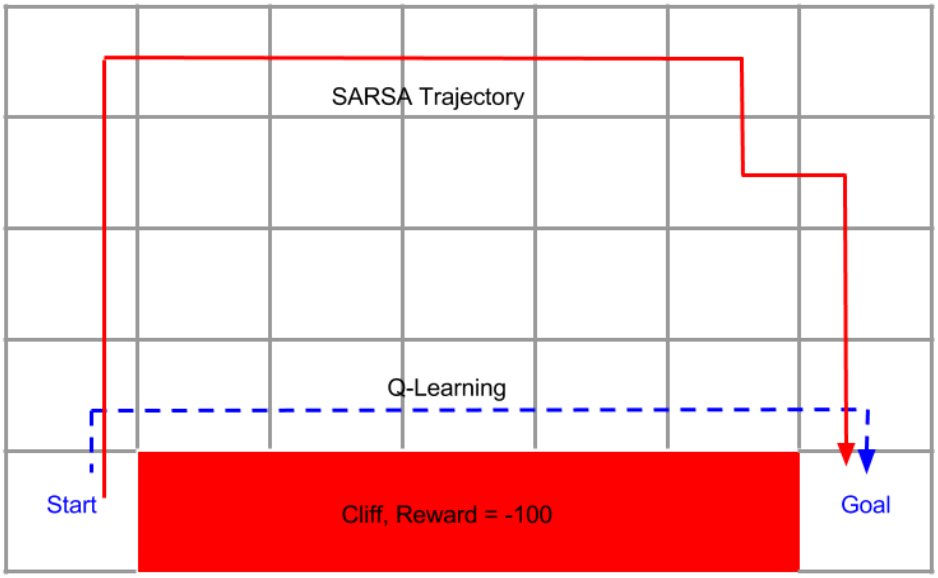
- **Grid size variation along the length**:
    - Grid Size [5X4]:



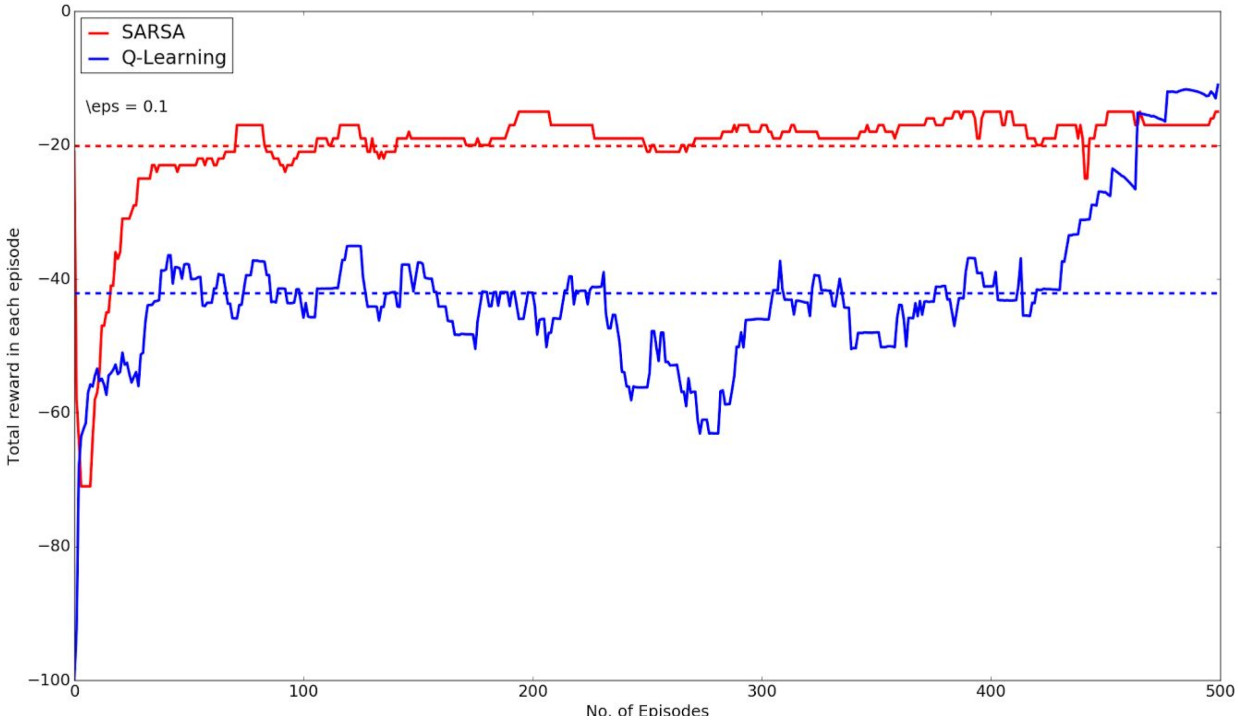


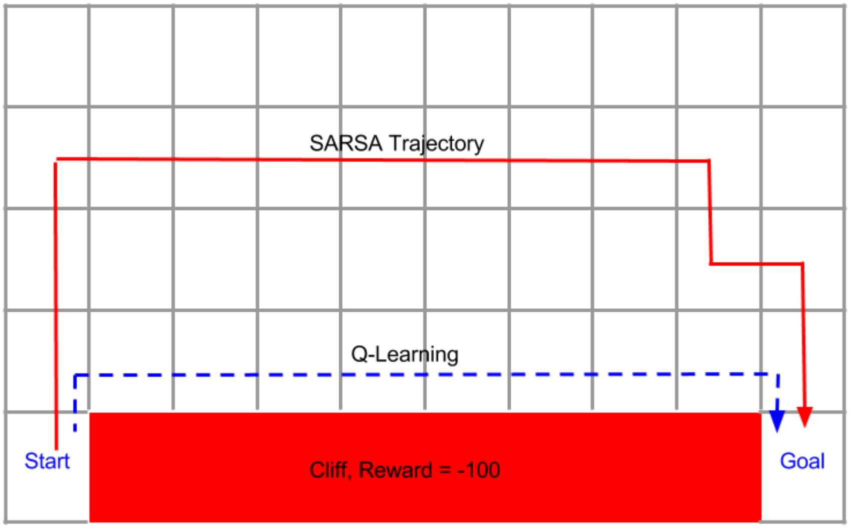
- Grid Size [5X7]:





- Grid Size [5X10]:

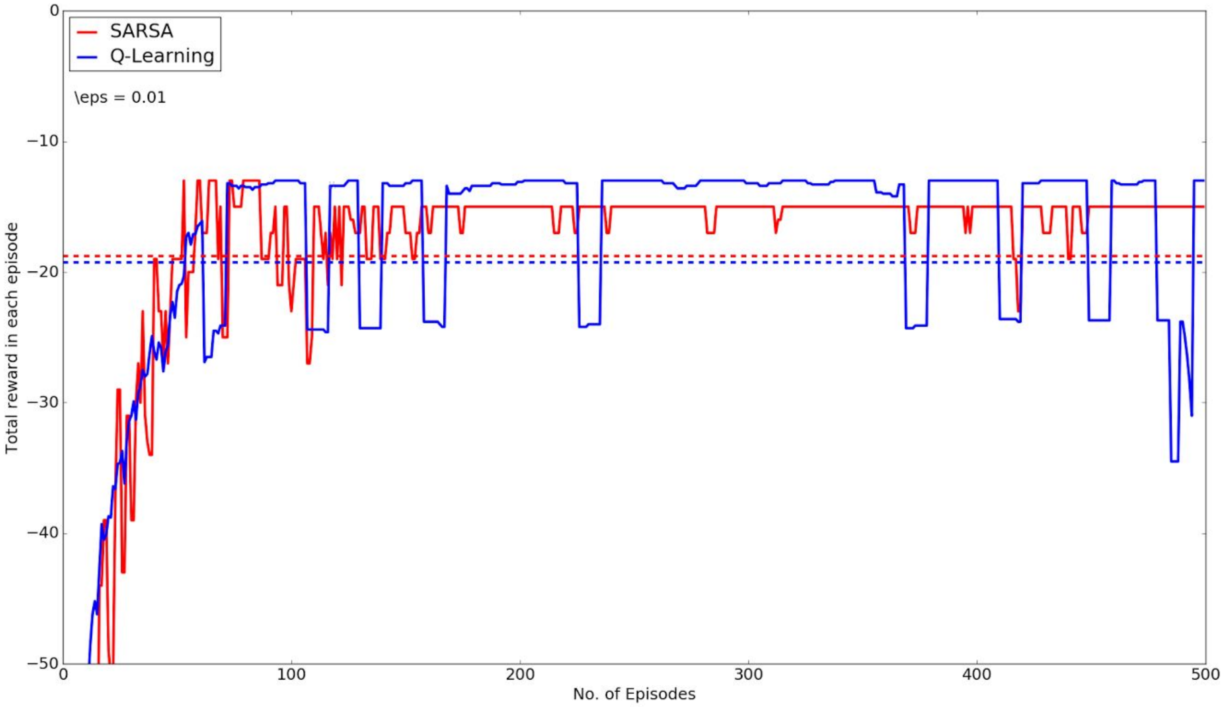


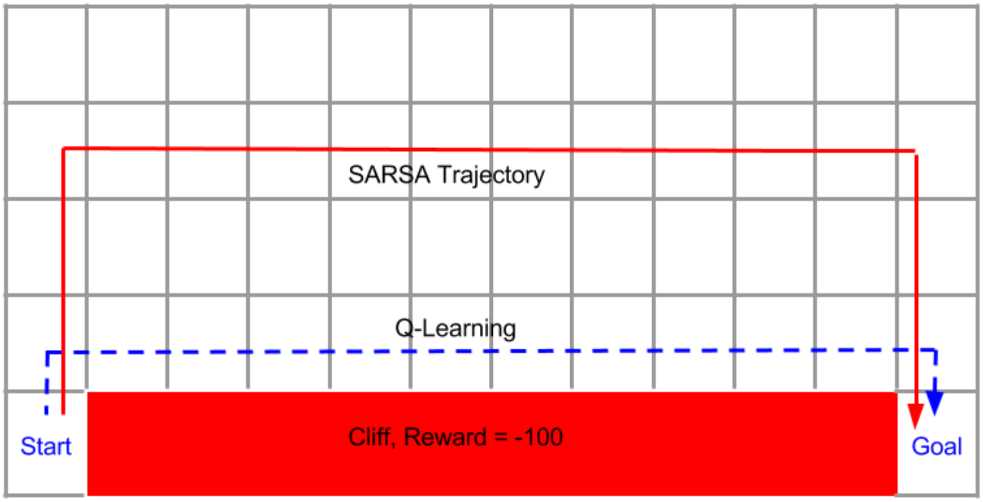


As the start state and the terminal state are symmetrical along the length of the grid world, hence it can be expected that *no significant change in the SARSA trajectory will be observed if the length of the grid world is varied*. The results validate this hypothesis.

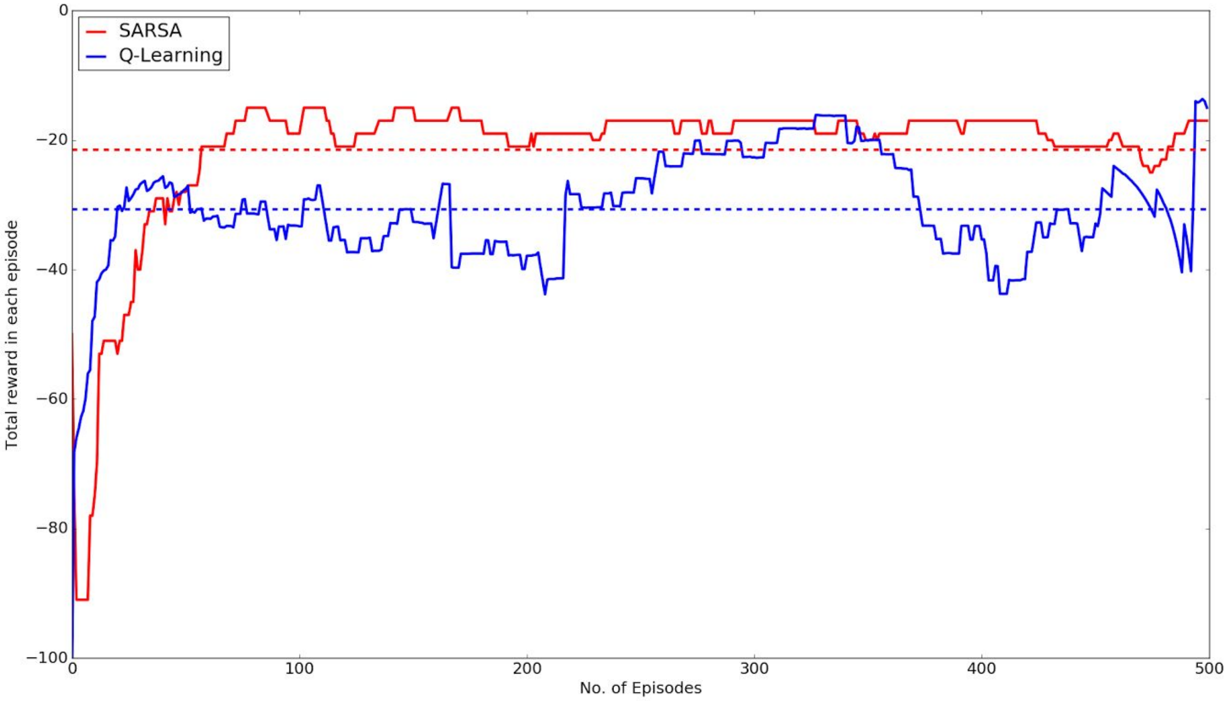
For analyzing the effects of the exploration factor, the grid world size was kept constant at [5X12] as based on the empirical evidence, the SARSA trajectory is independent of the size of the grid world.

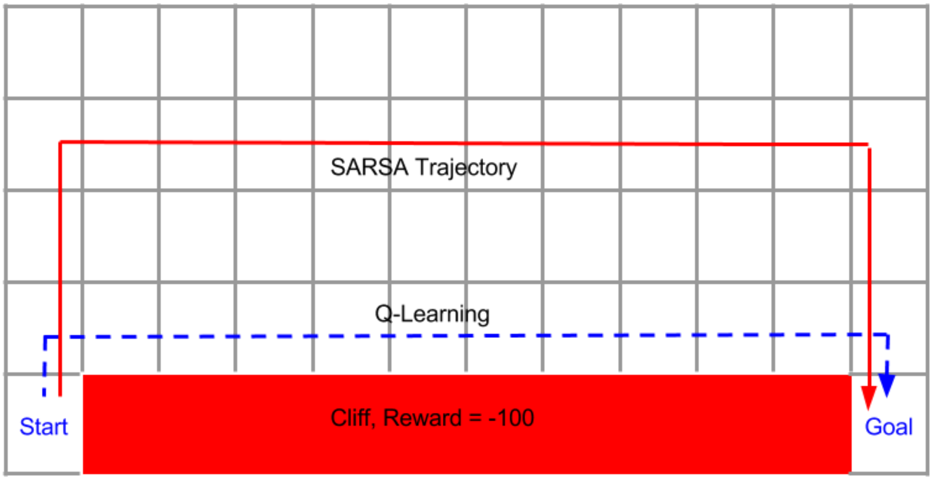
- **Effect of the exploration factor ε:**
  - ε = 0.01



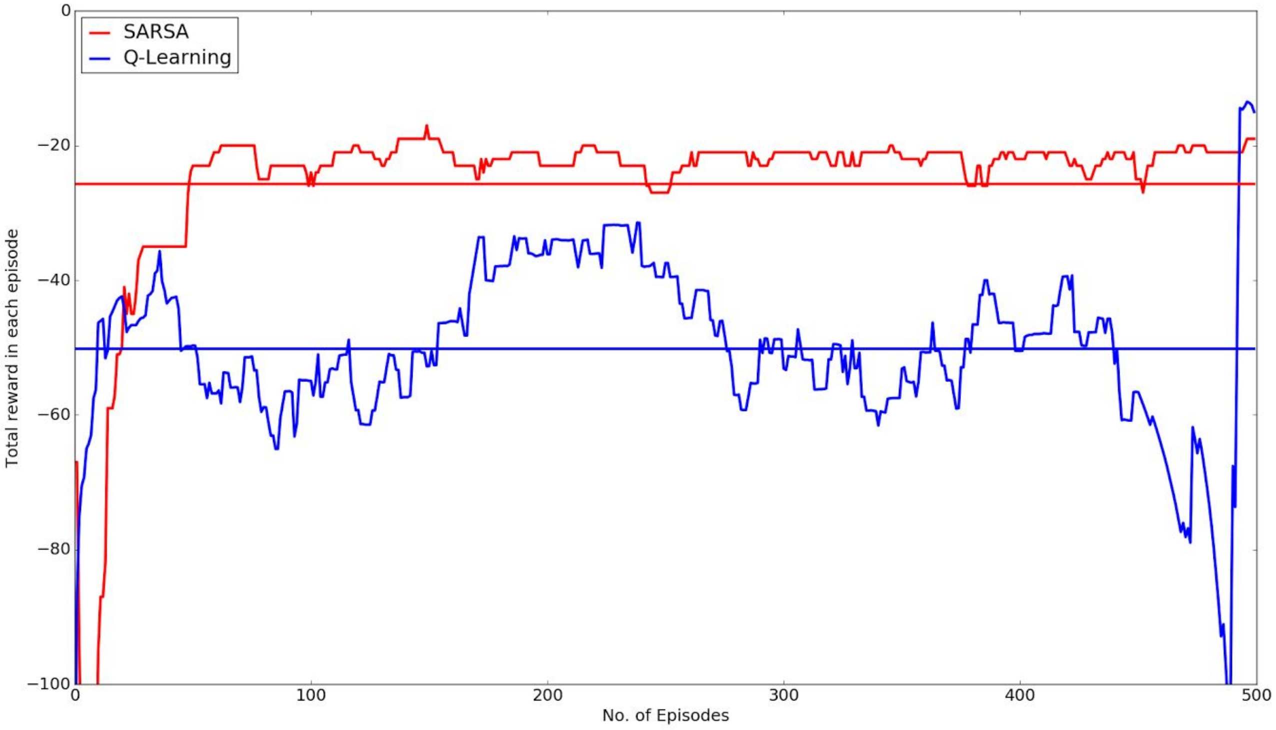


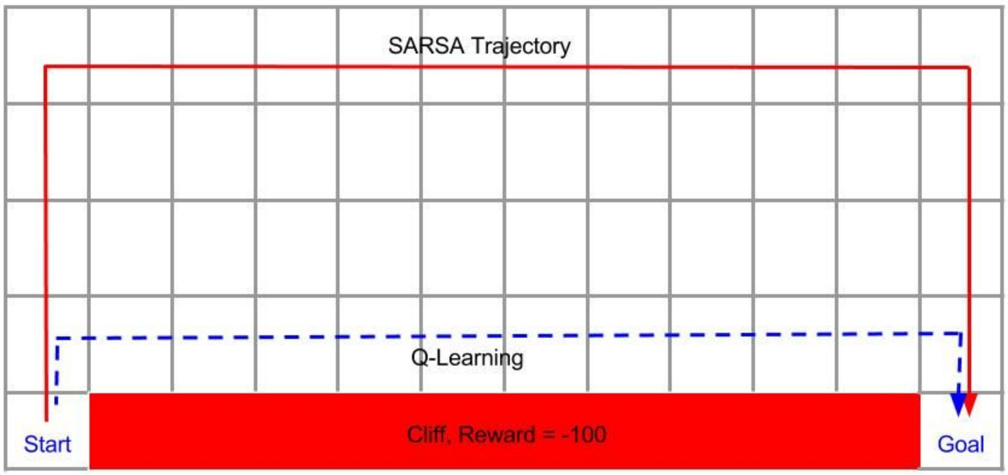
- ε = 0.05


SARSA
Q-Learning
Total reward in each episode
No. of Episodes
0
−20
−40
−60
−80
−100
0
100
200
300
400
500

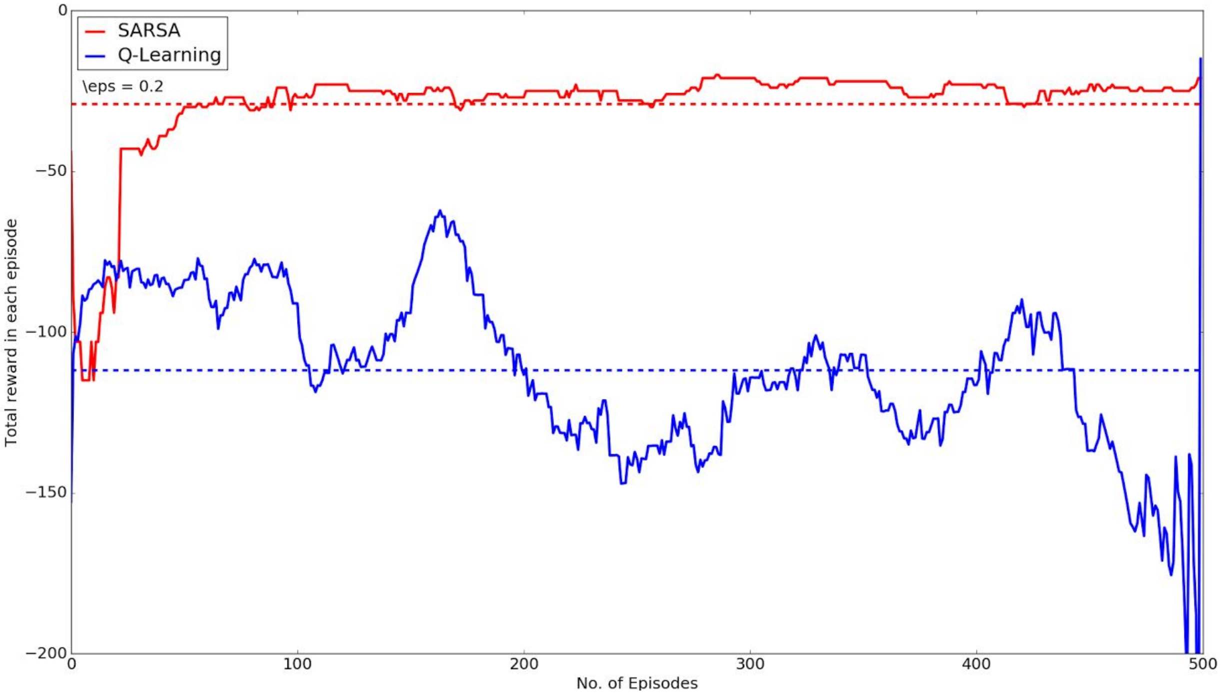

SARSA Trajectory
Q-Learning
Start
Cliff, Reward = -100
Goal

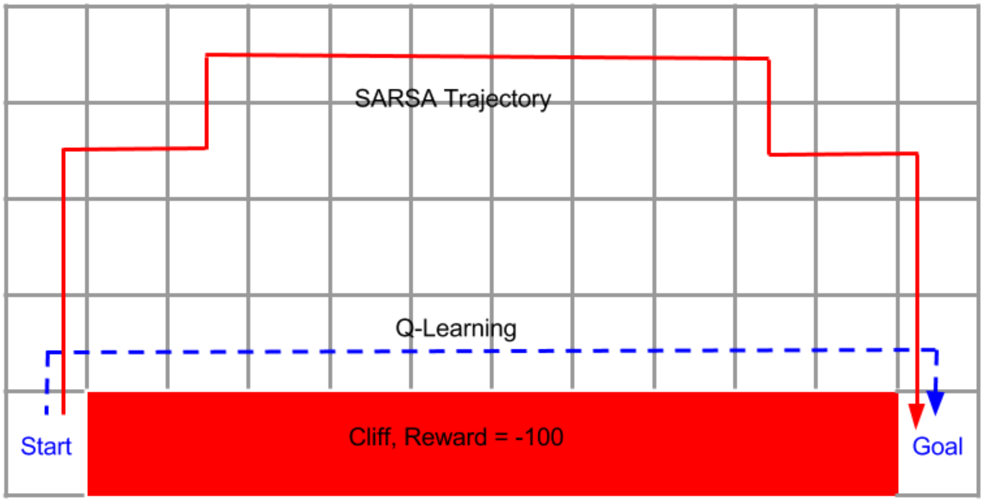
- ε = 0.1


SARSA
Q-Learning
Total reward in each episode
No. of Episodes
0
−20
−40
−60
−80
−100
0
100
200
300
400
500

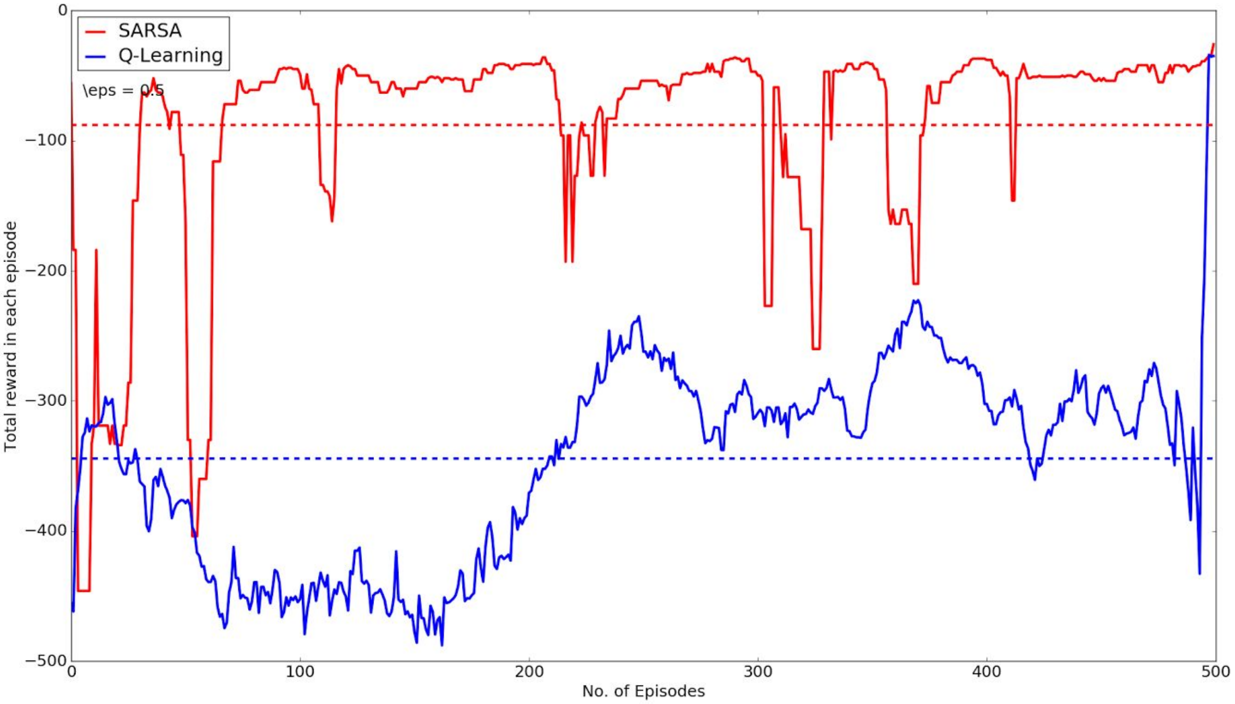

SARSA Trajectory
Q-Learning
Start
Cliff, Reward = -100
Goal

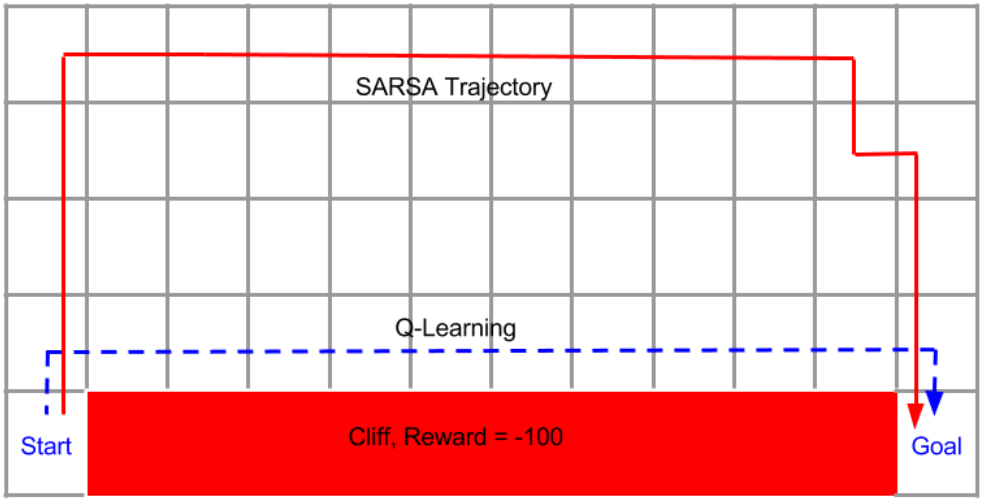
○ ε = 0.2


SARSA
Q-Learning
\eps = 0.2
Total reward in each episode
No. of Episodes
0
−50
−100
−150
−200
0
100
200
300
400
500


SARSA Trajectory
Q-Learning
Start
Cliff, Reward = -100
Goal

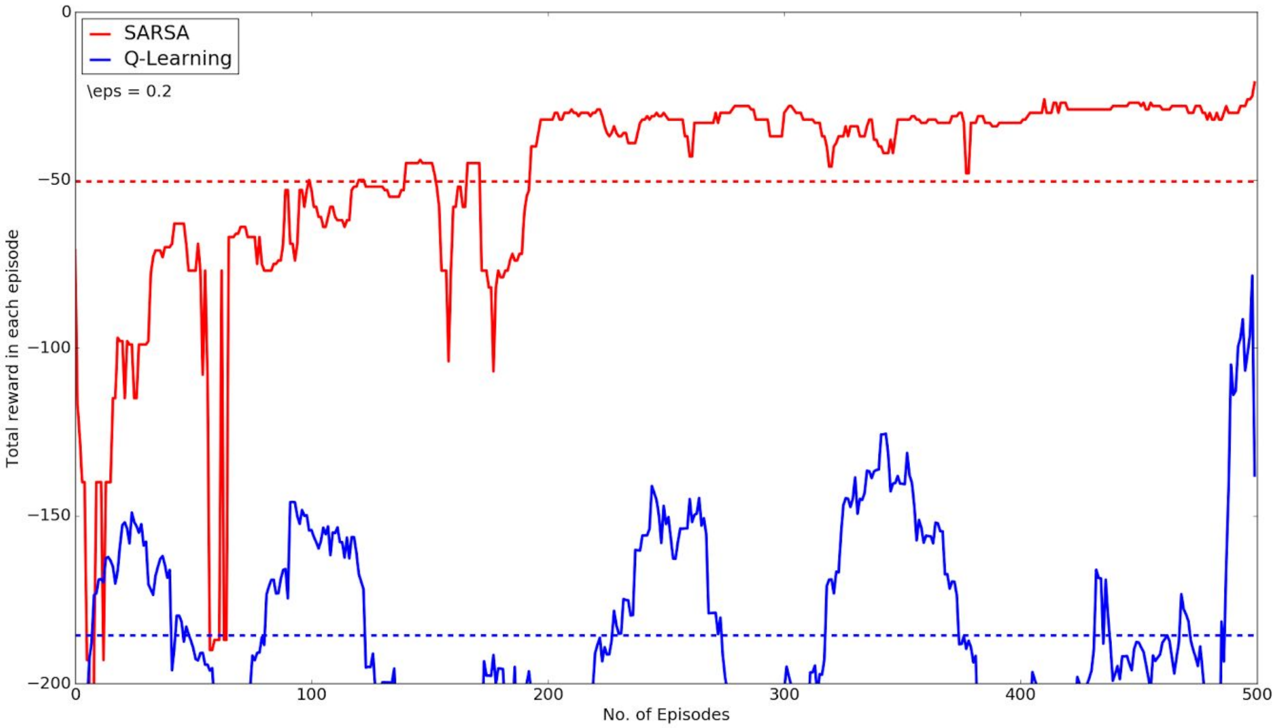
- ε = 0.5

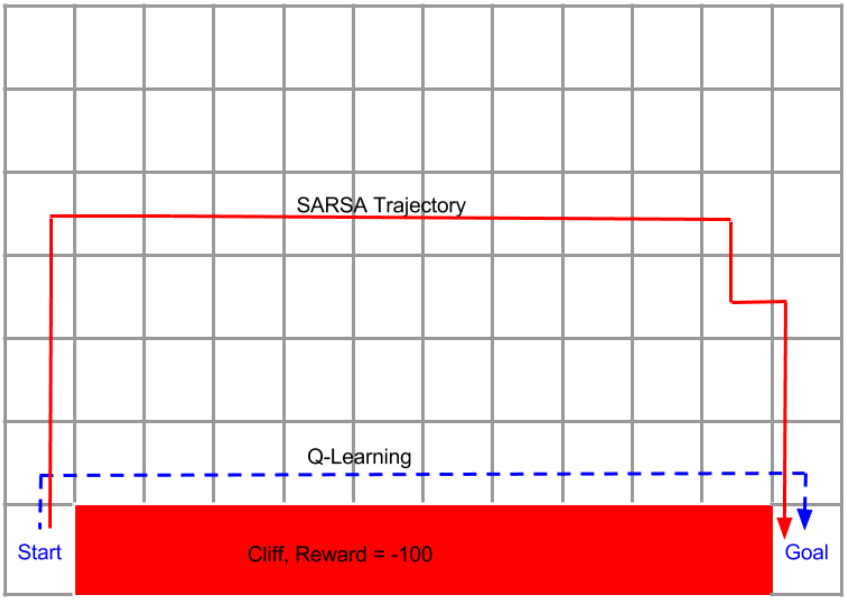




As can be seen from the results, with an increase in the exploration factor ε from 0.01 to 0.1, the algorithm tends to take more safer paths and increase its offset from the cliff. For higher exploration factors $\varepsilon = 0.2$ and ε= 0.5, there is no major shift in the SARSA trajectory in comparison to the ε =0.1 . This can be explained as the effect of increasing exploration factor diminishes after a critical offset distance from the cliff (a representative of path safety), because the second part of the reward function (conditioned over time) start to dominate in deciding the states with higher q-values.

→ Results to test the performance of the algorithms in grid world of size [7X12] with high exploration factor ε = 0.5





Comparison between off policy learning Q-Learning and the On-policy Learning SARSA:

As Q-learning is an off-policy learning method, hence the optimal trajectory learned by it does not get influenced by the changes in the exploration factor and the size of the grid world. However , In the case of the total returns, the off-policy learning method Q-learning goes down sharply as the exploration factor is increased. It can be understood as the chances to fall in the cliff by following the Q-Learning trajectory increases as the exploration factor is increased. As SARSA follows a “safe” trajectory, the reduction in the total returns with the increase in exploration factor is not as evident as in Q-Learning.

Conclusion:

In the present study, it was concluded that the trajectory taken by the Q-Learning algorithm is independent of the variation in the exploration factor. However, the total returns obtained during the episodes decreases as the exploration factor is increased. In the case of SARSA algorithm, as the algorithm is an on-policy learning algorithm, it tends to take more safer paths as the exploration factor is increased. But this effect diminishes after certain safety (distance from the cliff) is ensured. No significant effect of changing the size of the grid world was observed in the performance of the SARSA algorithm. Hence, the results validates the proposed hypothesis.